UNIVERSITY OF NOTRE DAME

Integrating Text into Psychological and Education Research

Latent Variable Modeling and Applications

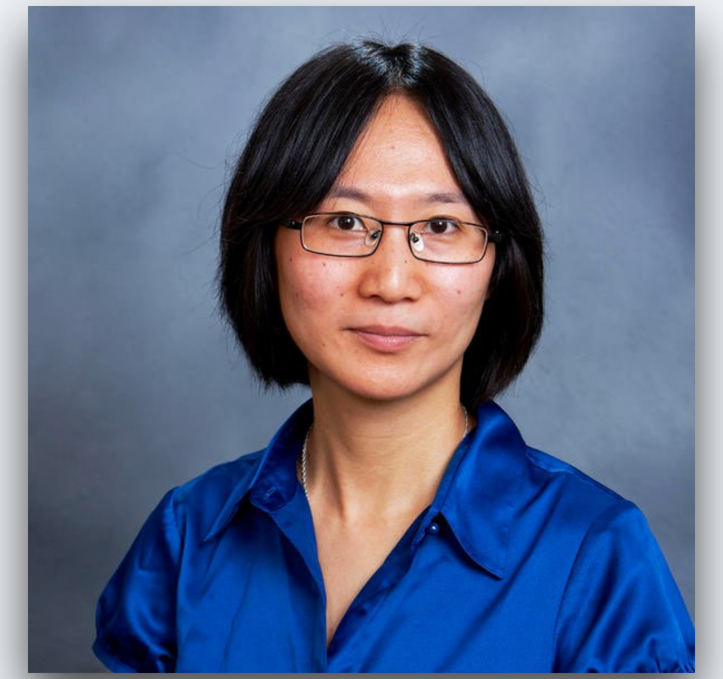
Kenneth Tyler Wilcox

Department of Psychology
University of Notre Dame

13 September 2021

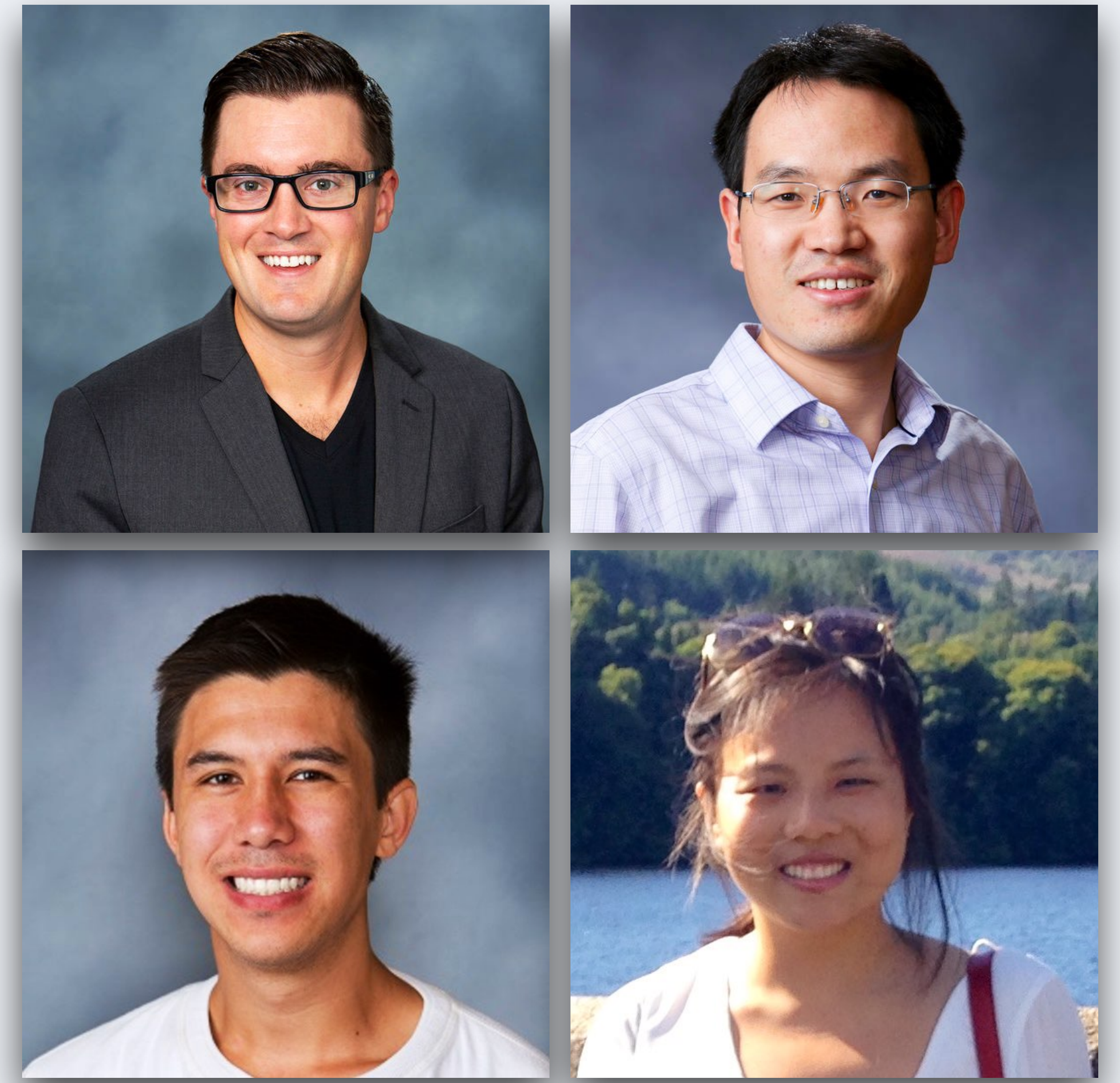
My Research

Cumulative Data Analysis



Applications

Text Mining



Outline

- Text data in psychology and education
- Dictionary methods
- Latent variable models
- A new model: supervised topic modeling with covariates
    - Estimation, interpretation, and software
    - Simulation study
    - Application to emotional dysregulation
- Future directions

Text Data in Psychology and Education

Text Data in Psychology and Education

- Long history in psychological research and educational assessment
  - Freud (1901)
  - General inquirer system (1966)
  - Linguistic Inquiry and Word Count (LIWC)
  - Topic modeling (2003)
  - Word embeddings

- Some applications
  - Measure student ability
  - Measure emotion
  - Study relationships
  - Early detection of depression
  - Identify prognostic risk factors for dementia
  - ...

(see, e.g., Bennet, 1991; Danner et al., 2001; Tausczik & Pennebaker, 2010)

But Why Not Scales?

What Are We Missing?

- Greater nuance in assessment
- Measure auxiliary or complementary information
- Closed-ended items may overemphasize testing skills, not construct domain
- Better measurement reliability
- Integration of qualitative and quantitative methods

(Boyle & Hutchinson, 2009; Ercikan et al., 1998; Jodoin, 2003; Kjell et al., 2018; Yang et al., 2018)

The Case of Two Participants

What Are We Missing?

- Data from study of nonsuicidal self-injury (NSSI) and emotional dysregulation (DERS)

- Px 1: NSSI = “yes”, Self-Rating = 7
  - DERS = 108

- Px 2: NSSI = “yes”, Self-Rating = 7
  - DERS = 63

What Are We Missing?

Interpersonal Conflict Narratives

- Px 1: NSSI = “yes”, Self-Rating = 7
  - DERS = 108
    - “Hanging out with **roommate** and **best friend**… **friend** cracked a joke that felt very insulting”

- Px 2: NSSI = “yes”, Self-Rating = 7
  - DERS = 63
    - “**Roommates** had **friends** over… they left a mess and never cleaned it in the kitchen”

Measurement: Dictionaries

Dictionary Methods

- LIWC is popular in social science research
  - Sentiment analysis
- **Predefine** constructs with lists of words

friend	joke	insulting	mess	...
2	1	1	0	...
1	0	0	1	...

Limitations of Dictionary Methods

- Inadequate scope of dictionary constructs
- Limited relevance
- Time-consuming and expensive to create
- Cannot account for polysemy

(Garten et al., 2018; Pennebaker et al., 2003)

Invariance and Polysemy

CRAWL

- Rate of movement
- Swimming stroke
- Moving on hands and knees
- “make your skin crawl”

Measurement: Latent Variable Models

Latent Semantic Indexing

(Not Really a Latent Variable Model)

- Effectively PCA for word frequencies
- Singular value decomposition of document-term matrix
- Eigenvectors and loadings interpreted as “semantic space”
- Over-fits the sample

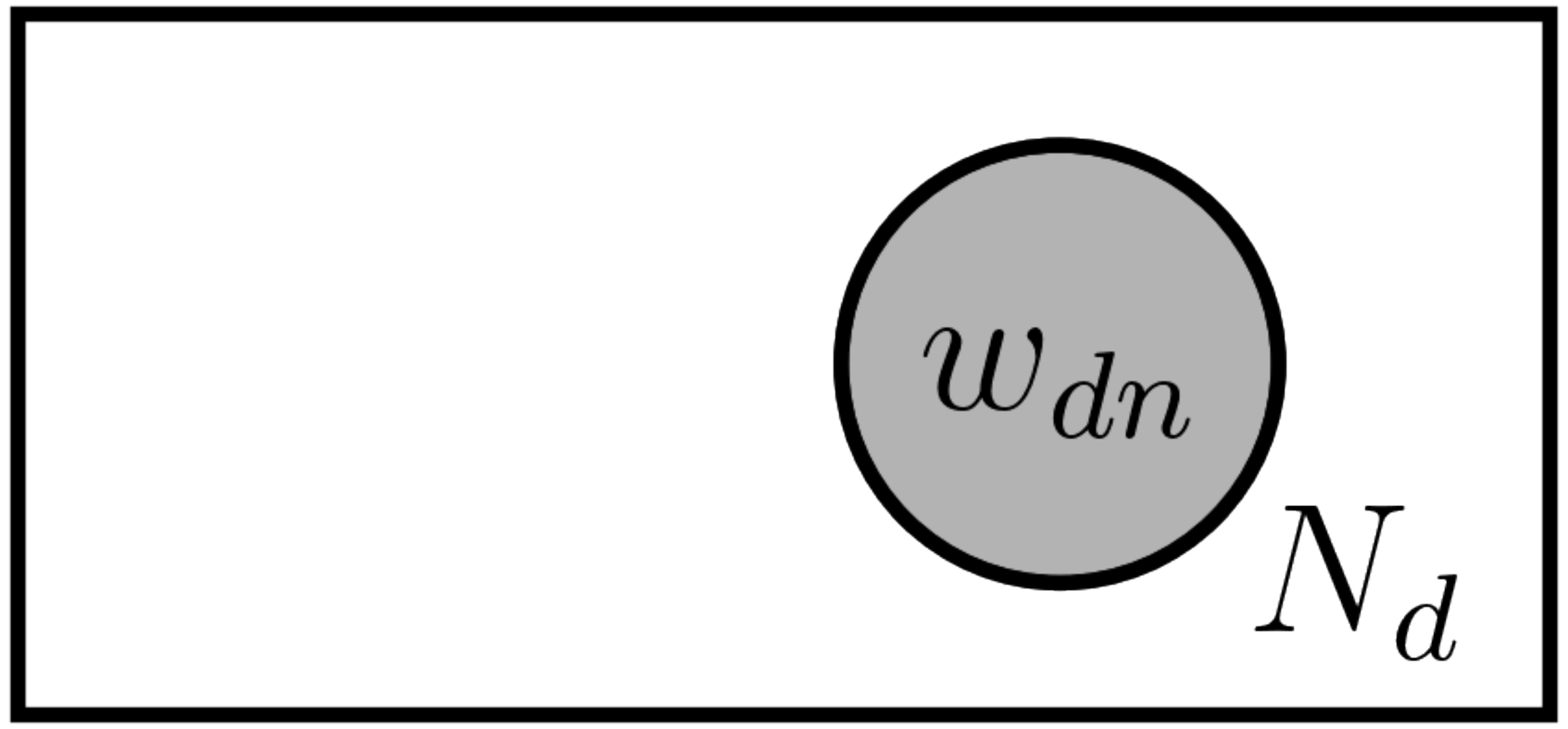
(Deerwester et al., 1990; Blei et al., 2003)

Topic Models

Latent Dirichlet Allocation

Probability distributions on words

Word w_{dn} for word 1, 2, ..., N_d



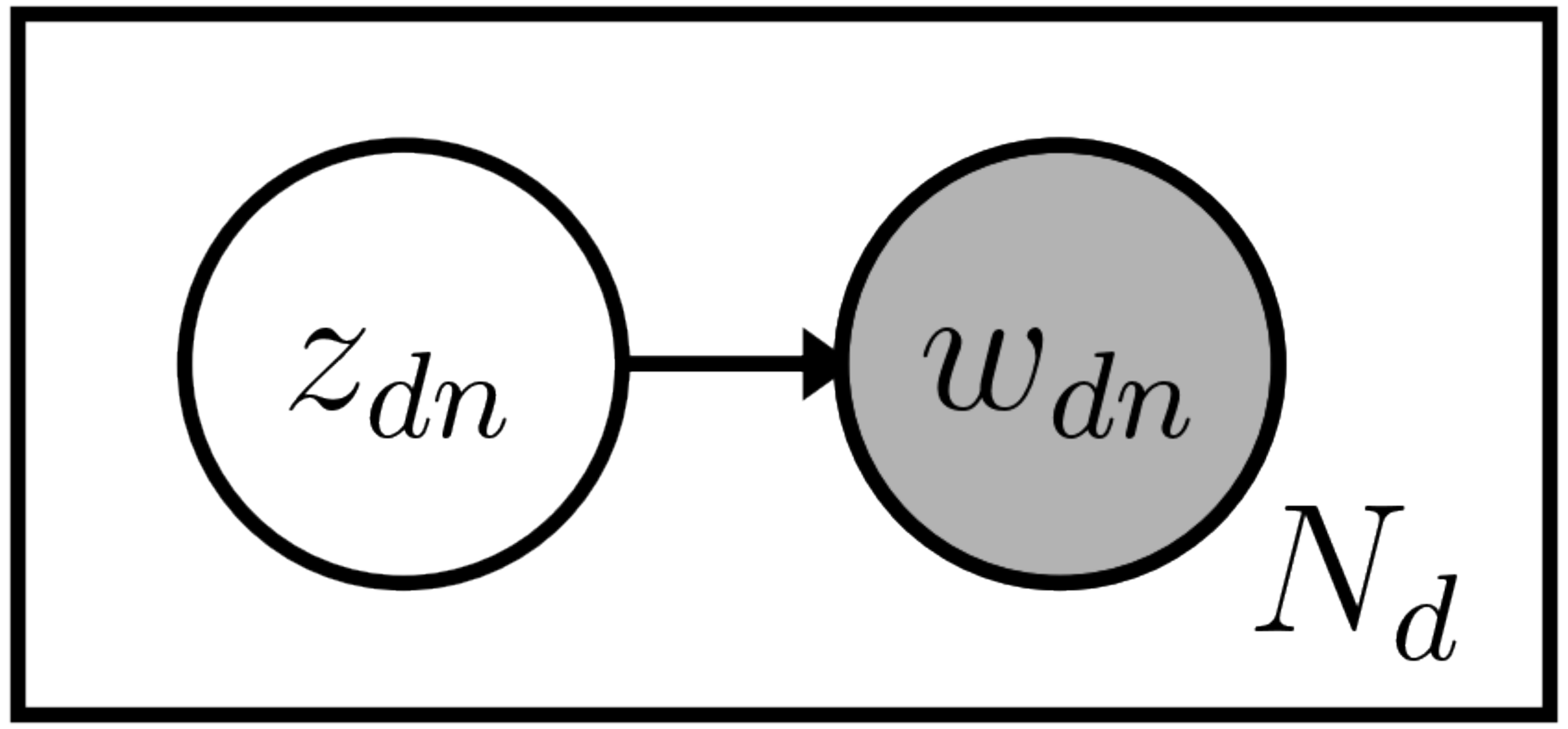
(Blei et al., 2003)

Topic Models

Latent Dirichlet Allocation

Probability distributions on words

Topic assignment z_{dn} for each word w_{dn}



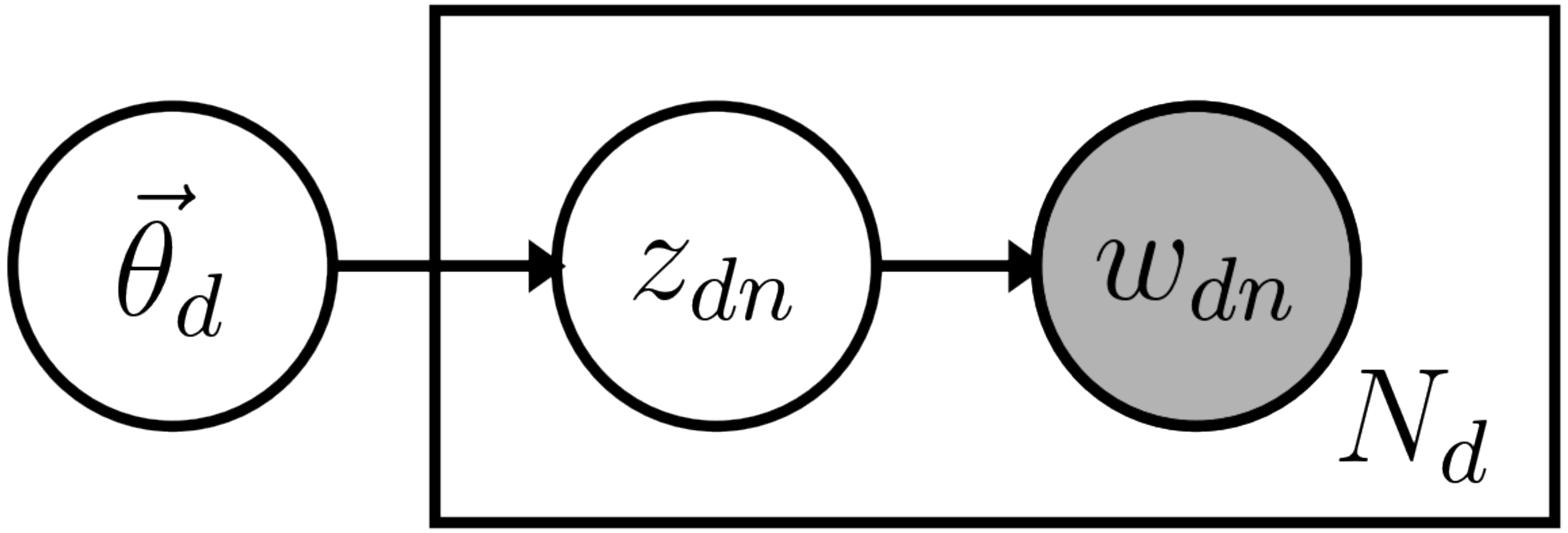
(Blei et al., 2003)

Topic Models

Latent Dirichlet Allocation

Probability distributions on words

Topic proportions $\overrightarrow{\theta}_d$ for each document



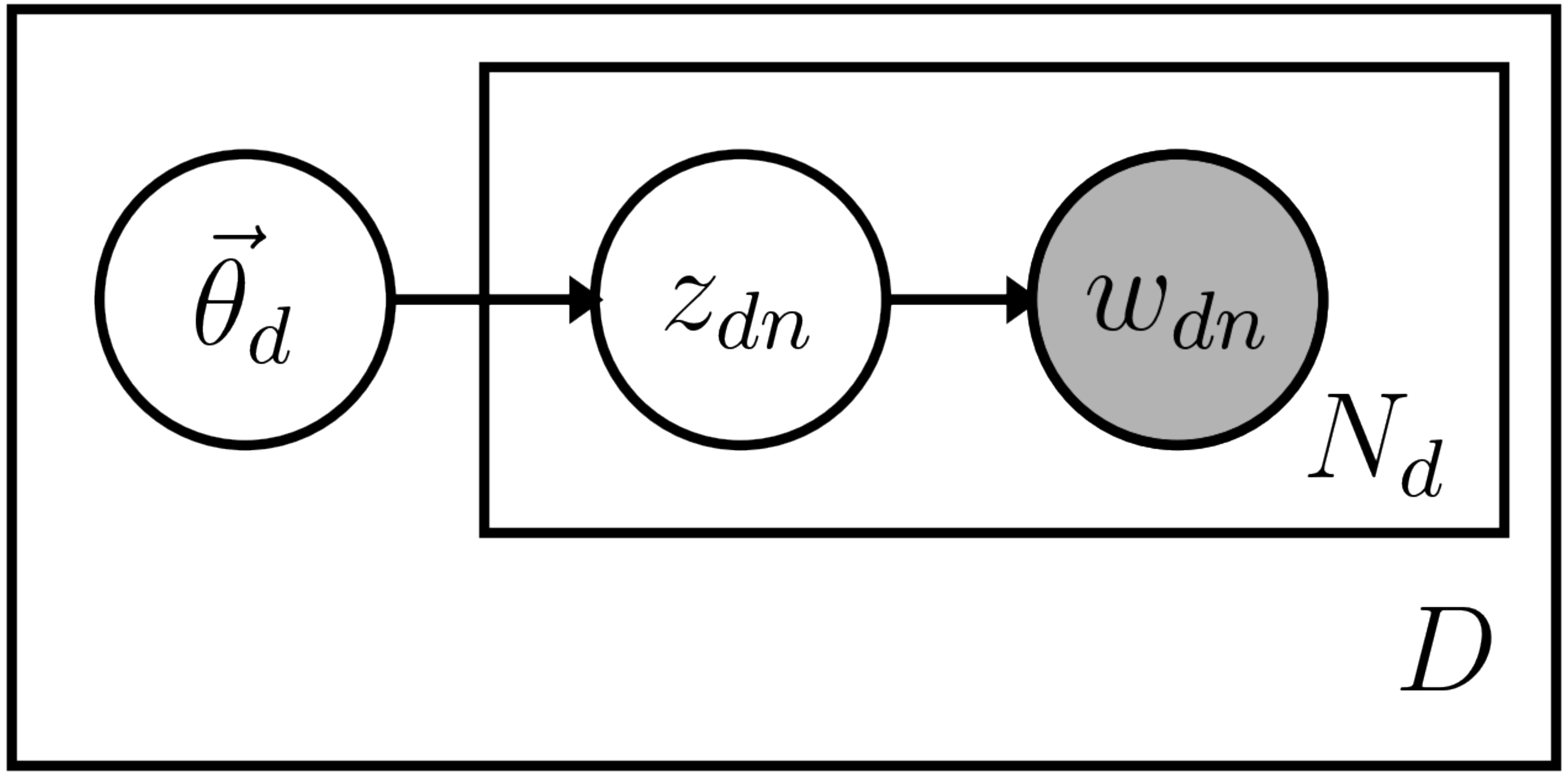
(Blei et al., 2003)

Topic Models

Latent Dirichlet Allocation

Probability distributions on words

Independent set of D documents



(Blei et al., 2003)

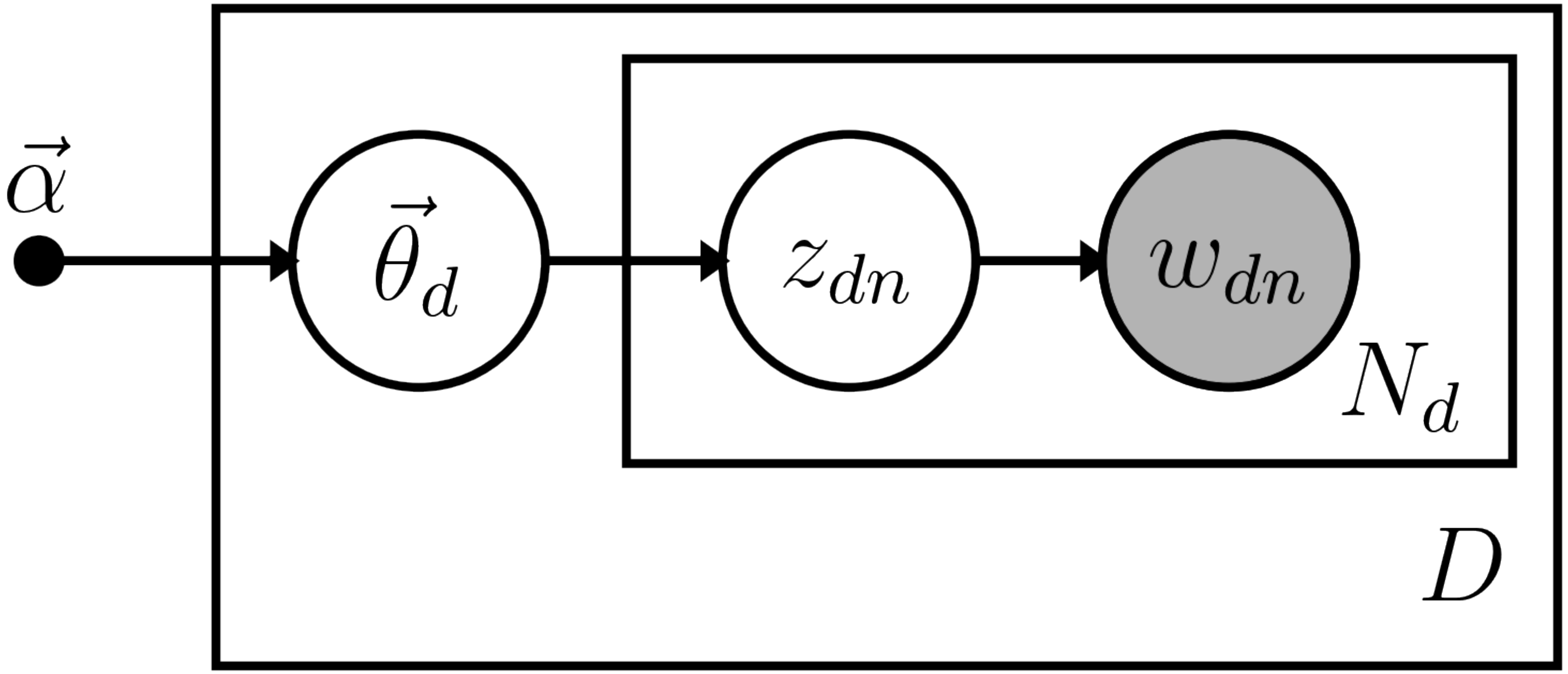
Topic Models

Latent Dirichlet Allocation

Probability distributions on words

Hyperparameter vector $\overrightarrow{\alpha}$ for topic proportions

Fixed across documents



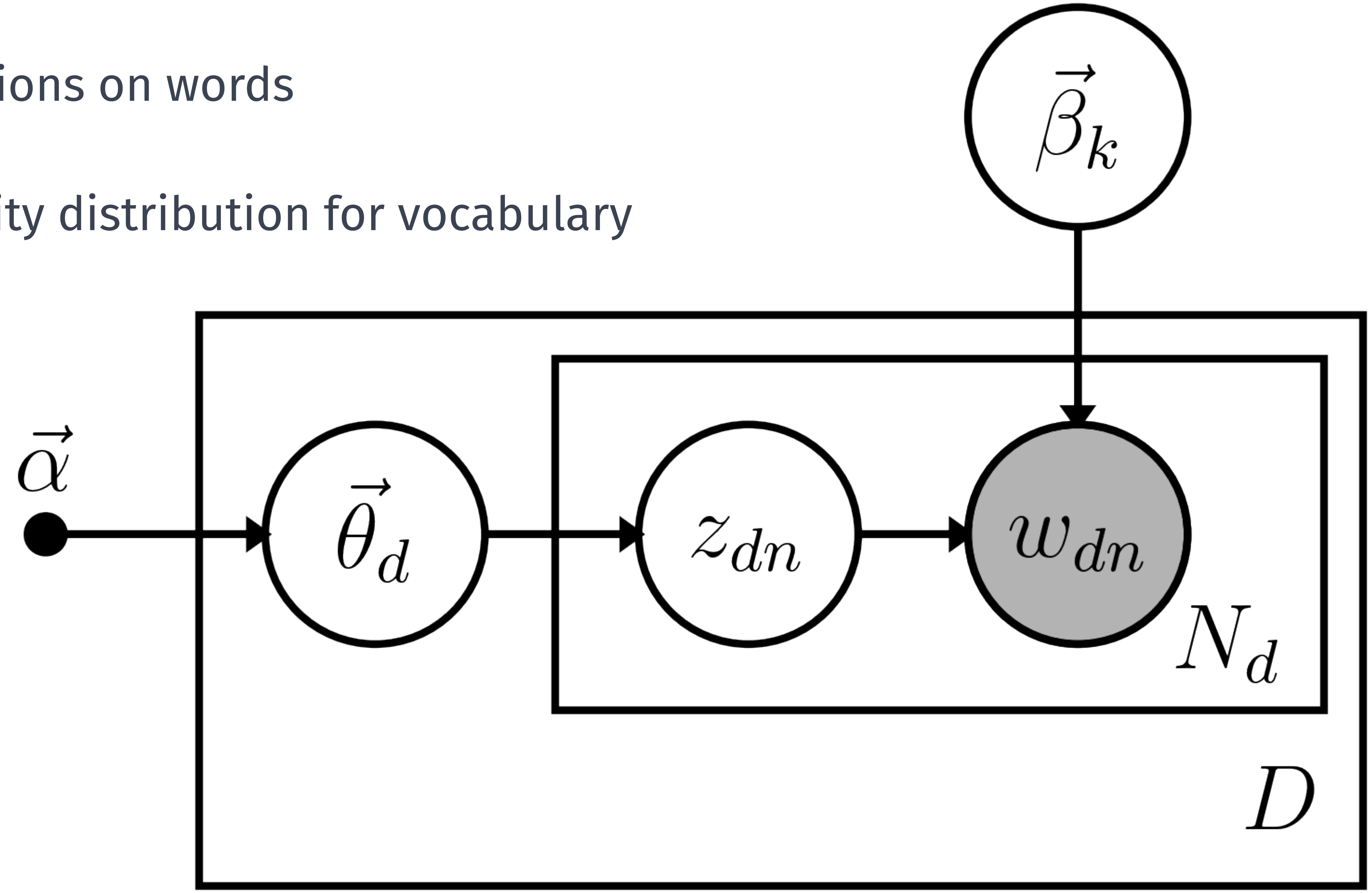
(Blei et al., 2003)

Topic Models

Latent Dirichlet Allocation

Probability distributions on words

Topic $\overrightarrow{\beta}_k$ = probability distribution for vocabulary



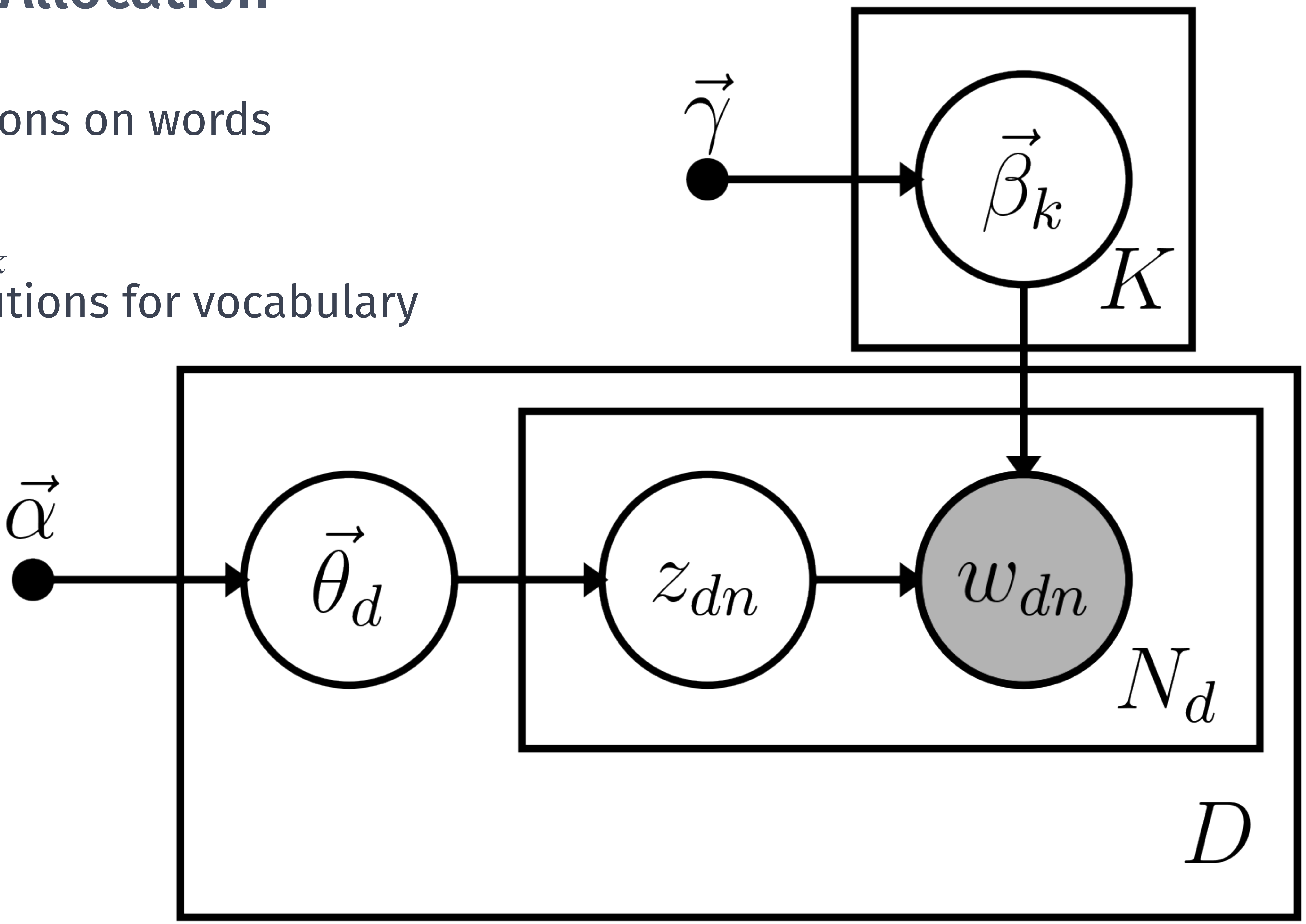
(Blei et al., 2003)

Topic Models

Latent Dirichlet Allocation

Probability distributions on words

K different **topics** $\overrightarrow{\beta}_k$
    K different distributions for vocabulary

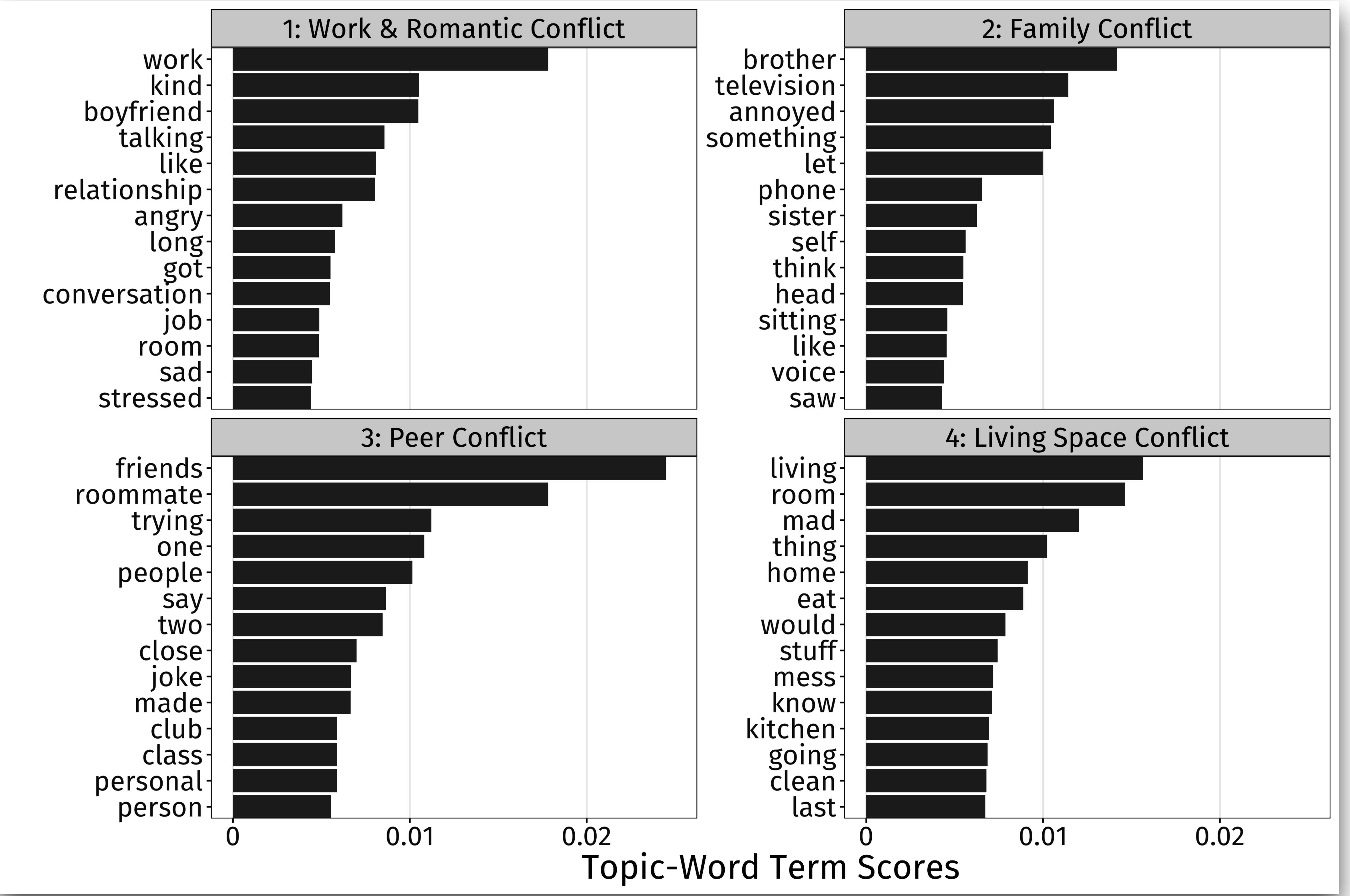


(Blei et al., 2003)

Illustration: Topics

Interpersonal Conflict Narratives

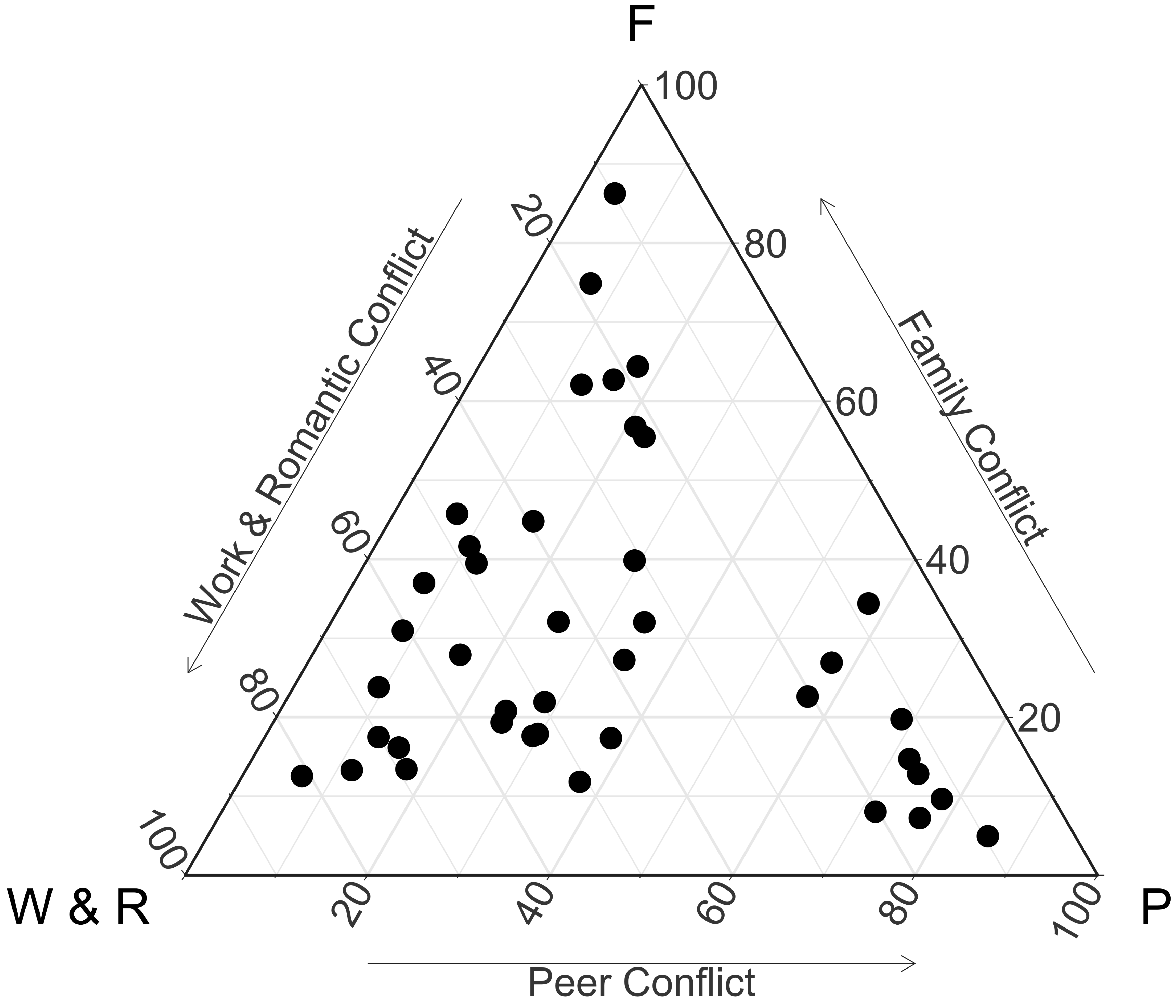
- Topic 1: Work & Romantic Conflict
  - “dad came to visit her at work… embarrassed and angry…”
  - “she and boyfriend had argument about being in a long distance relationship… she wants to move… he has to stay for his job”
- Topic 2: Family Conflict
  - “mom and dad just got a divorce… argument with mom and brother…”
- Topic 3: Peer Conflict
  - “friend made joke about her body in class… a little sad and hurt…”
- Topic 4: Living Space Conflict
  - “ex-roommate trashed the house and she was p***ed”



Wilcox, Jacobucci, Zhang, & Ammerman (under review)

Illustration: Topic Proportions

Conditioning on Living Space Conflict



Putting Topics in Context

- Like latent factors, researchers have linked topics to other measures

$Y =$  $\eta + X\beta + \epsilon$

- Surprisingly, an appropriate model is unavailable

(e.g., Finch et al, 2018; He, 2013; Kim et al., 2017; Rohrer et al., 2017)

Regression with Topics

Current Practice

- Two-stage approach
  - 1. Estimate topic proportions
  - 2. Use topic proportion **estimates** as regression predictors
- Two-stage approaches with latent variable models are problematic
- Current interpretation and inferential procedures for topics are incorrect

(Bakk, Tekle, & Vermunt, 2013; Packard et al., 2020; Petersen et al., 2012; Rohrer et al., 2017; Vermunt, 2010; Hayes & Usami, 2020)

Supervised Topic Modeling with Covariates (SLDAX)

Wilcox, Jacobucci, Zhang, & Ammerman (under review)

Funding Acknowledgement: Data presented in this talk was supported by NIMH 1F31MH107156-01A1 awarded to Brooke A. Ammerman.

Research Objectives

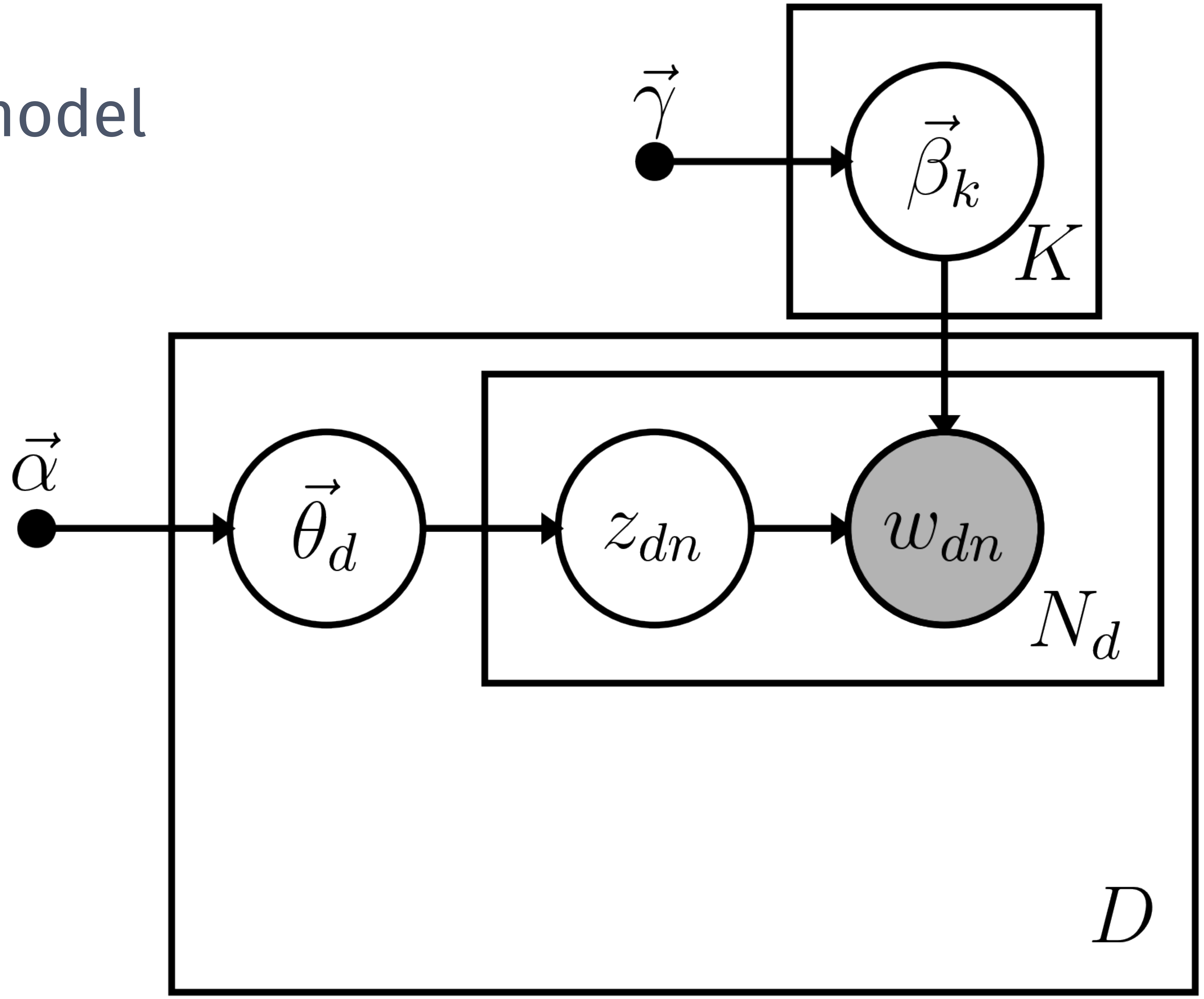
Wilcox, Jacobucci, Zhang, & Ammerman (under review)

- Develop new model to include covariates and topics to predict an outcome
- Evaluate estimation accuracy and efficiency of two-stage approach and our model
- Propose method to yield interpretable topic effects and correct inferences

Proposed Model

SLDAX — Supervised Latent Dirichlet Allocation with Covariates

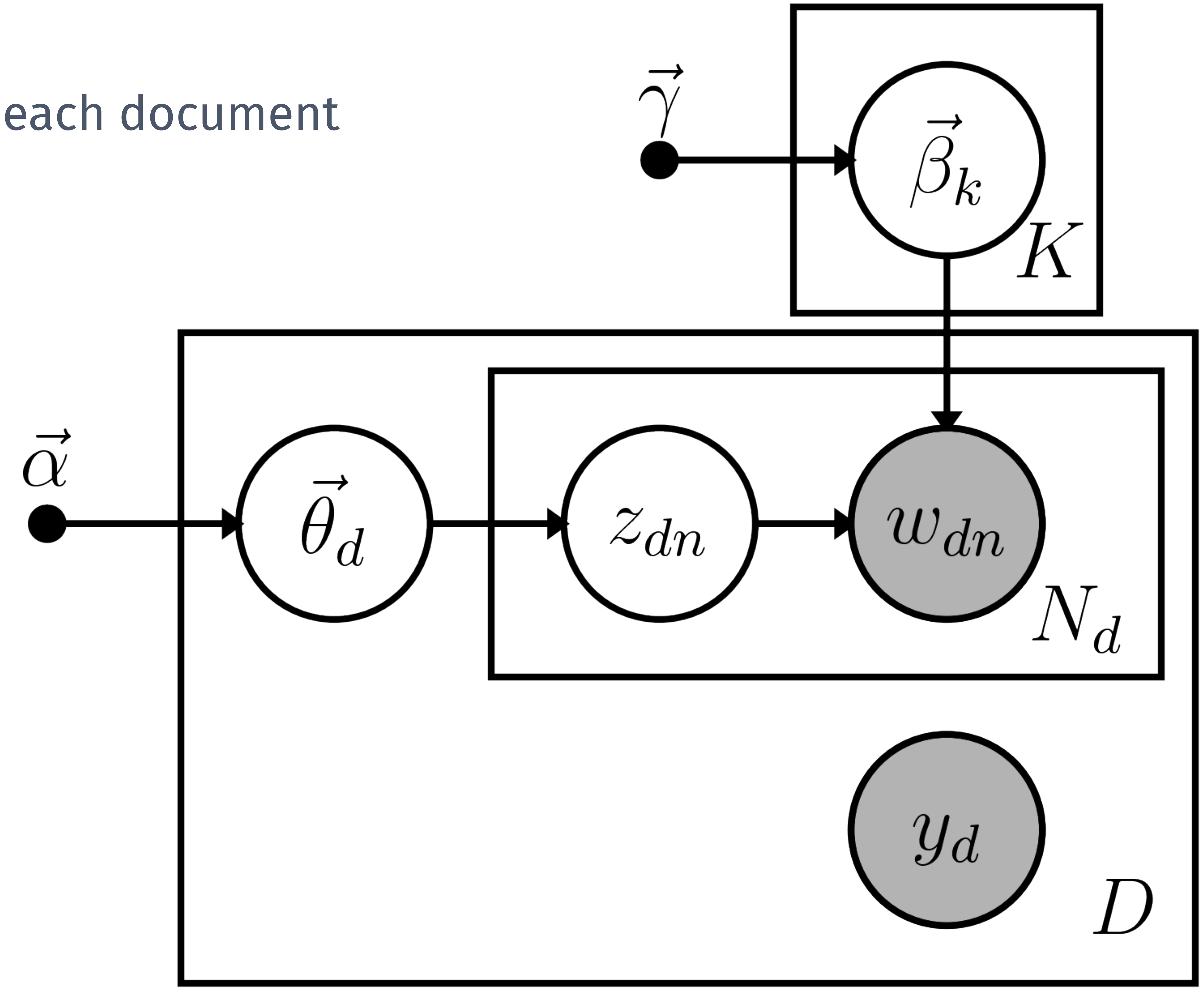
- Extend LDA model



Proposed Model

SLDAX — Supervised Latent Dirichlet Allocation with Covariates

- Outcome y_d with each document

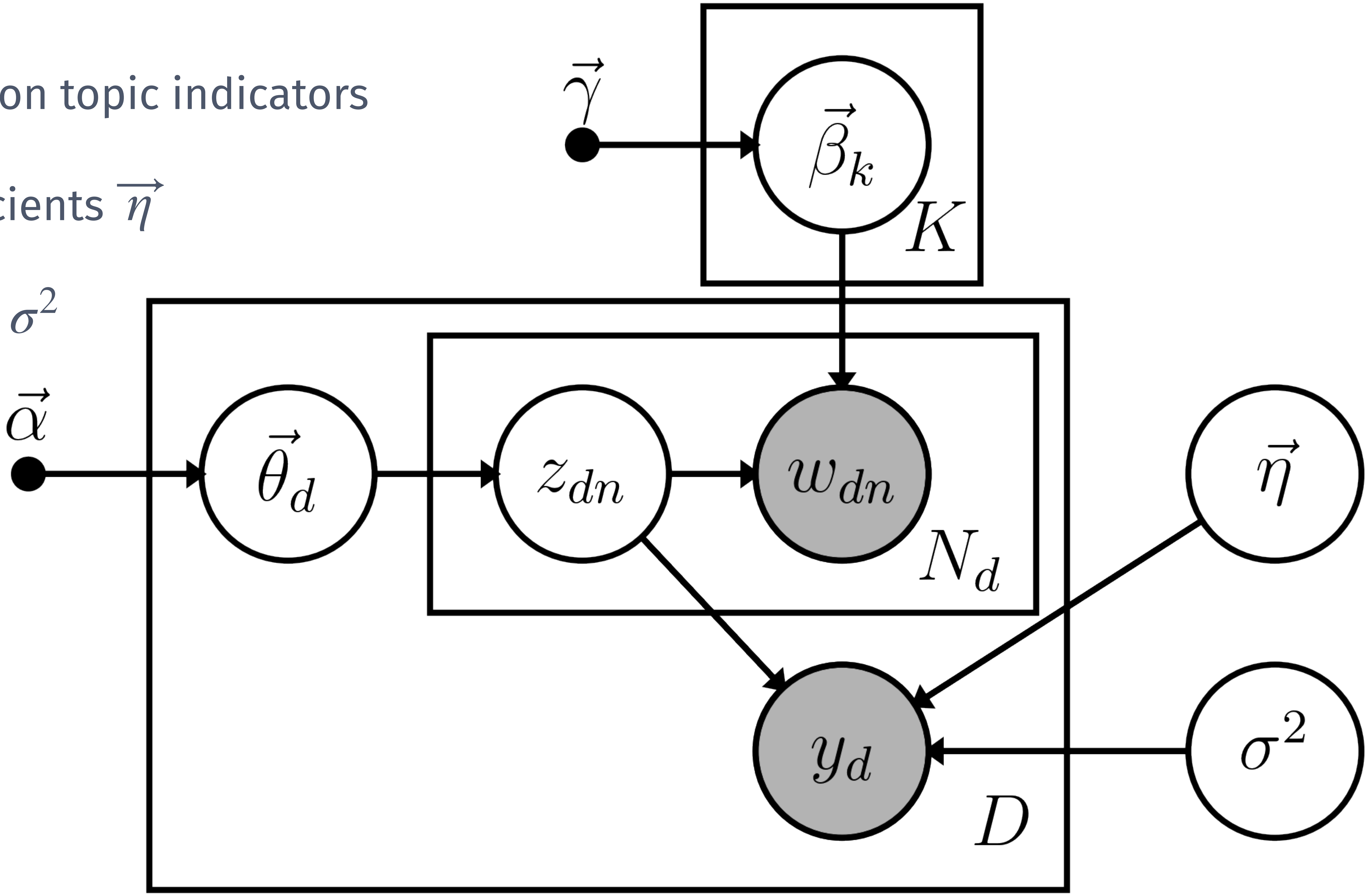


(Blei et al., 2010)

Proposed Model

SLDAX — Supervised Latent Dirichlet Allocation with Covariates

- Regress outcome on topic indicators
- Regression coefficients $\overrightarrow{\eta}$
- Residual variance σ^2

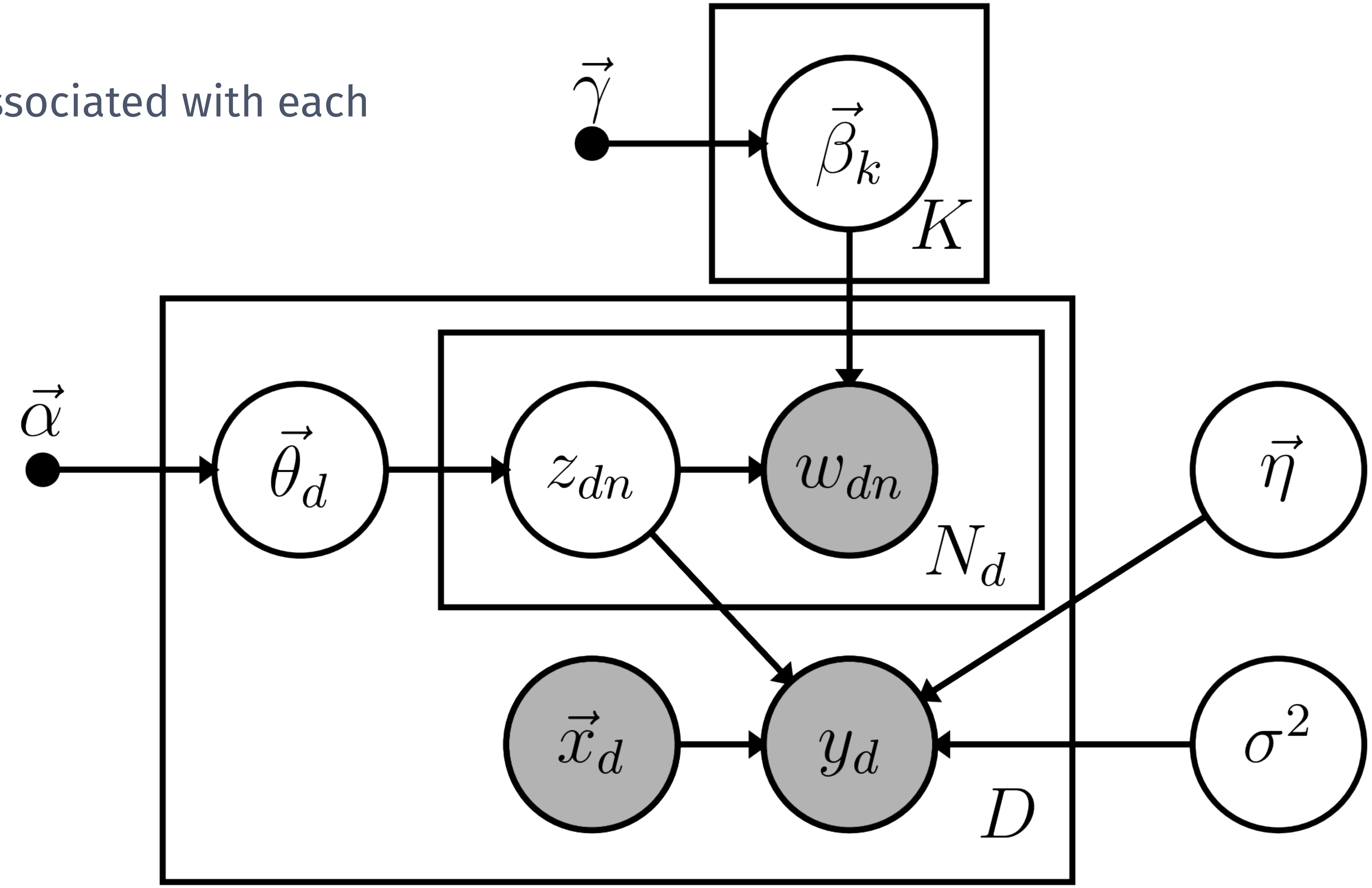


(Blei et al., 2010)

Proposed Model

SLDAX — Supervised Latent Dirichlet Allocation with Covariates

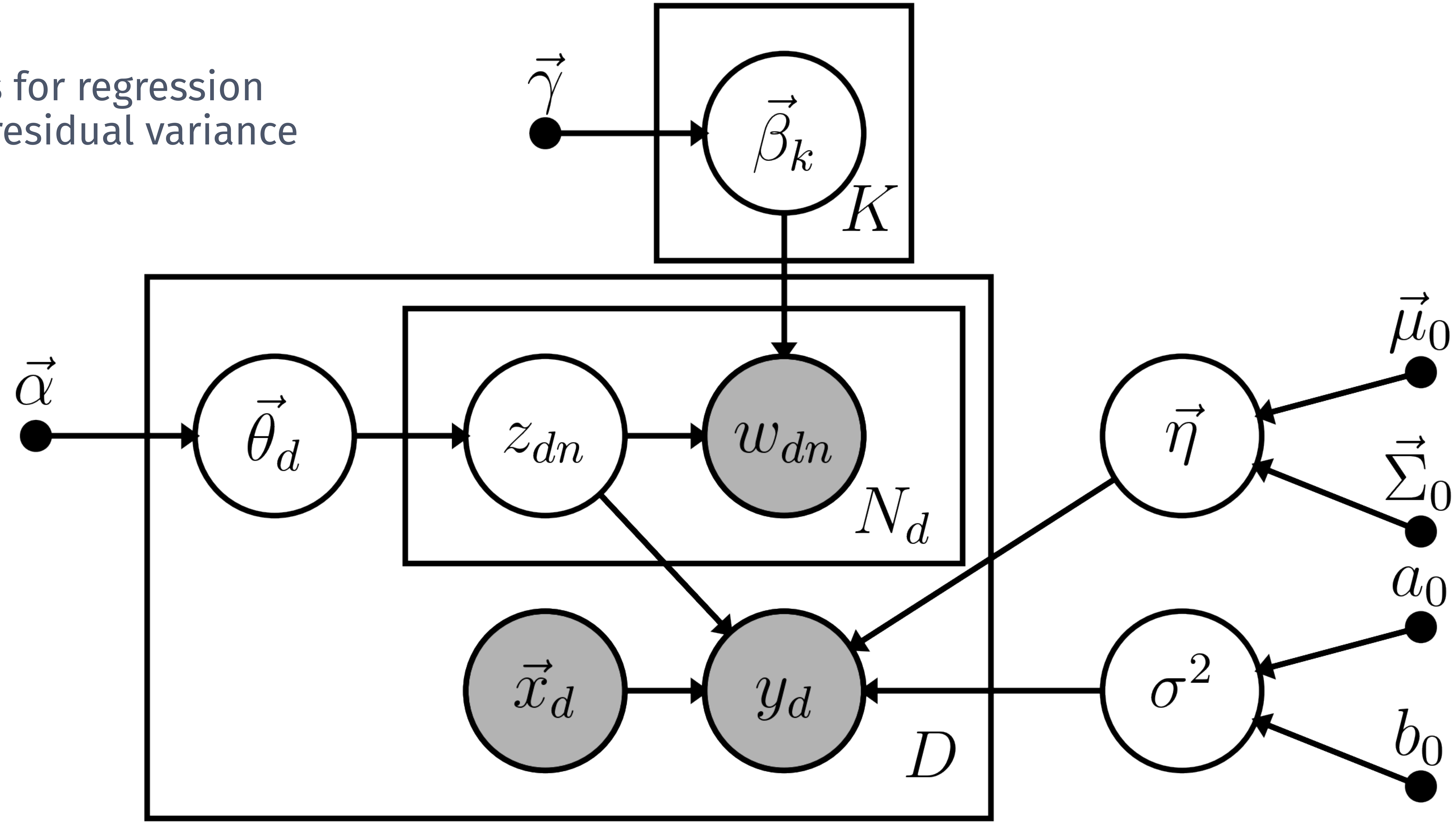
- Covariates $\vec{x}_d$ associated with each document



Proposed Model

SLDAX — Supervised Latent Dirichlet Allocation with Covariates

- Hyperparameters for regression coefficients and residual variance



SLDAX Model

$$\mathbb{E}\left[Y_d | \vec{X}_d, \vec{\bar{Z}}_d\right] = \sum_{k=1}^{K} \eta_k \bar{Z}_{dk} + \sum_{j=1}^{p} \eta_j X_{dj}$$

- Extended by generalized linear model framework to normal and dichotomous outcomes
- (Collapsed) Gibbs/Metropolis sampler for Bayesian estimation
  - Speed up mixing, reduce autocorrelation in chain
- Potential label switching handled by Stephens's algorithm

(Cassiday et al., 2020; Dias & Wedel, 2004; Liu, 1994; Stephens, 2000)

More in Paper

Wilcox, K. T., Jacobucci, R., Zhang, Z., & Ammerman, B. A. (under review). Supervised latent Dirichlet allocation with covariates: A Bayesian structural and measurement model of text and covariates. *PsyArXiv*. doi: 10.31234/osf.io/62tc3

$$L\left(\vec{\Theta}, \vec{B}, \vec{\eta}, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{D}{2}} \exp\left\{-\left(2\sigma^2\right)^{-1} \sum_{d=1}^{D} \left(y_d - \vec{r}_d'\vec{\eta}\right)^2\right\} \prod_{d=1}^{D} \prod_{n=1}^{N_d} \theta_{dz_{dn}} \beta_{z_{dn} w_{dn}}$$

$$f\left(\vec{\eta}, \sigma^2, \vec{\Theta}, \vec{B}, \vec{z}_1, \ldots, \vec{z}_D | \vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D\right) = \frac{L\left(\vec{\Theta}, \vec{B}, \vec{\eta}, \sigma^2\right) f\left(\vec{\eta}\right) f\left(\sigma^2\right) \prod_{d=1}^{D} f\left(\vec{\theta}_d\right) \prod_{k=1}^{K} f\left(\vec{\beta}_k\right)}{f\left(\vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D\right)}$$

$$\vec{\eta}| \cdot \sim \mathrm{N}\left(\vec{\eta}_1, \vec{\Sigma}_1\right)$$

$$\vec{\Sigma}_1 = \left(\vec{\Sigma}_0^{-1} + \vec{R}'\vec{R}(\sigma^2)^{-1}\right)^{-1} \qquad \vec{\eta}_1 = \vec{\Sigma}_1\left(\vec{\Sigma}_0^{-1}\vec{\mu}_0 + \vec{R}'\vec{y}(\sigma^2)^{-1}\right)$$

$$\sigma^2| \cdot \sim \mathrm{IG}\left(\frac{a_0 + D}{2}, \frac{1}{2}\left[b_0 + \left(\vec{y} - \vec{R}\vec{\eta}\right)'\left(\vec{y} - \vec{R}\vec{\eta}\right)\right]\right)$$

$$f\left(z_{dn} = k|\cdot\right) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(y_d - \vec{r}_d'\vec{\eta}\right)^2\right\} \times \frac{\left(n_{w_{dn}k}^{(-dn)} + \gamma\right)\left(n_{dk}^{(-dn)} + \alpha\right)}{n_k^{(-dn)} + V\gamma}$$

Software

R Package

- `psychtm`
  - In development
  - Estimation for topics models (LDA, supervised LDA, SLDAX)
  - Written in C++ for speed
- Available on Github

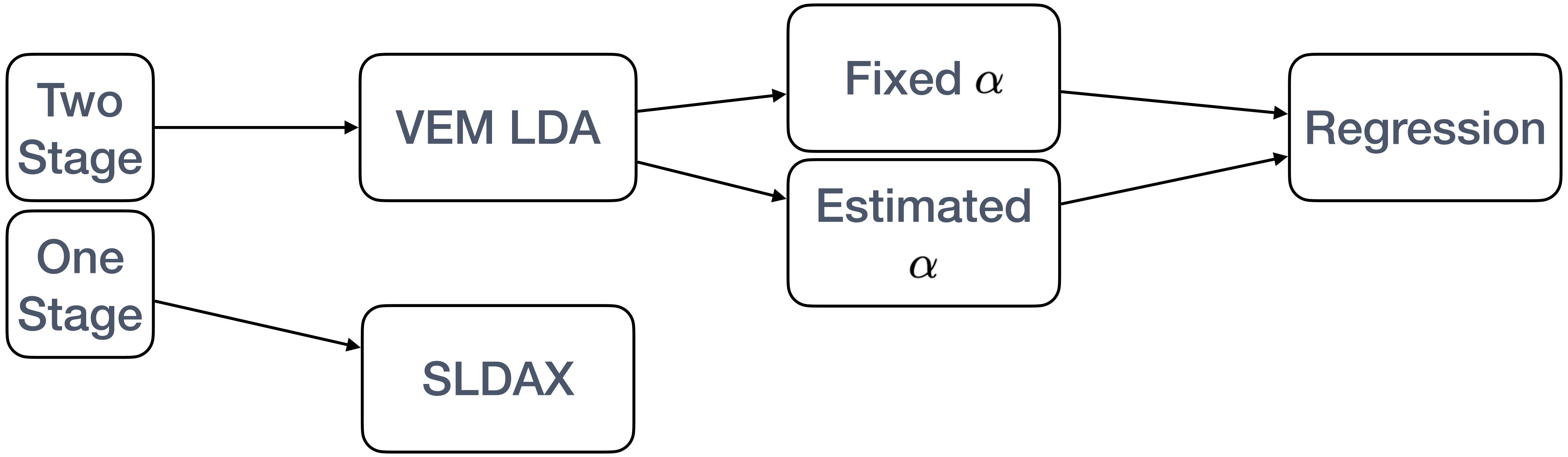
```
devtools::install_github("ktw5691/psychtm")

fit <- gibbs_sldax(y ~ x1 + x2, data = xy, docs = docs, K = 2, V = nvocab)
```

Simulation Study

Design and Methods

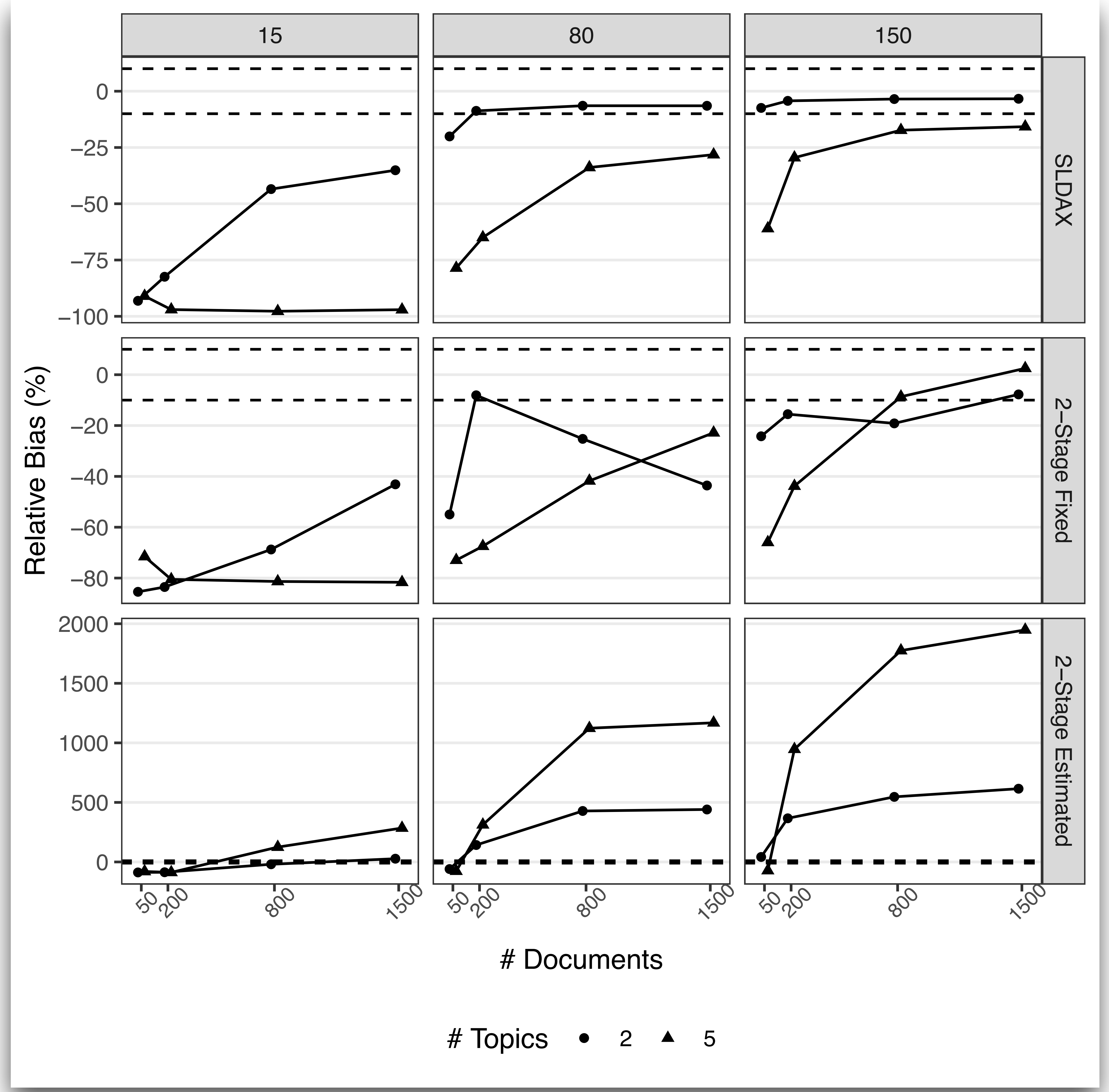
- Key Conditions
  - # topics: {2, 5}
  - # subjects: {50, 200, 800, 1500}
  - Average document length: {15, 80, 150}
- Methods



Simulation Results

Two-Stage vs. SLDAX

- Two-stage method
  - **Overestimated** regression coefficients w/ estimated hyperparameter
    - This gets worse with more data!
  - Inconsistent (?) w/ fixed hyperparameter
- SLDAX estimates less biased
  - Require adequate sample size and document lengths
  - Can be underestimated
  - More efficient (smaller MSE) — not shown here



Interpretation & Inference

Topic Regression Coefficients

- Topic proportions are ipsative (i.e., sum to 1)
- Corresponding regression coefficients are conditional means of the outcome when **only** that topic is present
  - Common to see all positive or all negative coefficients
  - Meaning depends on conditional mean of outcome
  - Generally, cannot compare them to 0

Interpretation & Inference

Contrasts

- Define the "effect" of a topic on the outcome with contrasts, e.g.,

$$c_k = \eta_k - \frac{\sum_{k' \neq k}^{K} \eta_{k'}}{K-1} \stackrel{?}{=} 0$$

- Sample c_k from posterior distribution
- We can interpret the sign and credible interval w.r.t. 0
- Better weighting using Piepel's method

(Park, 1978; Piepel, 1982; Snee et al., 1976)

Empirical Application

Relationships Among Nonsuicidal Self-Injury and Interpersonal Stress

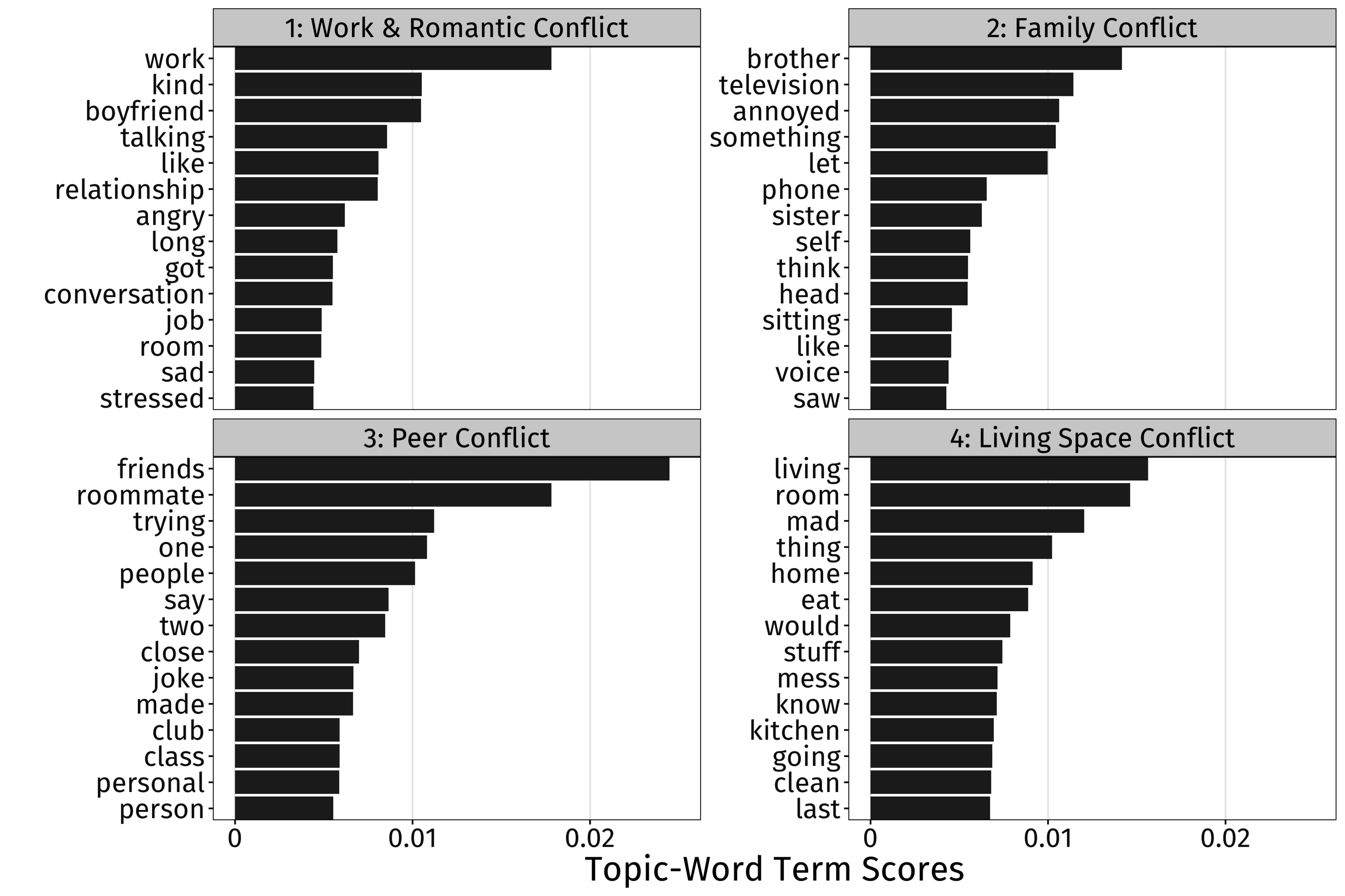
- Undergraduate sample (n = 41); majority (84%) identified as female
- 56% reported NSSI history
- Interview transcripts about a recent upsetting interpersonal interaction
  - After pre-processing, median word count = 63
- Self rating of degree of upset/distress for the interaction (Likert: 1—10)
- Modeled **emotional dysregulation (DERS)** with
  - NSSI history
  - Self rating
  - Interpersonal interaction narrative transcripts

(Gratz et al., 2011)

Empirical Application

Topics Measured by Interpersonal Interaction Interviews

- Topic 1: Work & Romantic Conflict
  - “dad came to visit her at work... embarrassed and angry...”
  - “she and boyfriend had argument about being in a long distance relationship... she wants to move... he has to stay for his job”
- Topic 2: Family Conflict
  - “mom and dad just got a divorce... argument with mom and brother...”
- Topic 3: Peer Conflict
  - “friend made joke about her body in class... a little sad and hurt...”
- Topic 4: Living Space Conflict
  - “ex-roommate trashed the house and she was p***ed”



Empirical Application

SLDAX Regression Results

- NSSI history associated with greater DERS
- Topics from negative interpersonal interaction jointly explain significant variability in DERS, $\Delta R^2 = 15\%$
  - NSSI and self rating explain 24%
  - Topic effects likely attenuated

	Coefficient / Contrast (SE)	95% BCI
Self Rating	0.7 (2.2)	[-3.7, 5.0]
NSSI History	21.7 (6.5)	[8.8, 34.4]
T1: Romantic & Work Conflict	92.3 (10.42) 9.7 (12.3)	[71.4, 112.6] [-15.1, 33.7]
T2: Family Conflict	67.0 (11.9) -20.4 (13.3)	[43.2, 90.7] [-46.3, 6.4]
T3: Peer Conflict	101.0 (10.3) 19.5 (11.6)	[80.7, 121.8] [-3.5, 42.3]
T4: Living Space Conflict	75.4 (11.5) -10.8 (13.1)	[52.1, 97.8] [-36.8, 14.9]

Summary

- Developed new model to incorporate topic model for text into regression framework
- Proposed model yields more accurate and efficient estimates than two-stage approach used in standard practice
- Document length is key for improving regression estimates
- Number of documents/subjects is key for power
- Contrasts are needed for interpretation and inference
- Text can measure what available scales may not

Future Directions

- Integrate topic model and IRT model for closed-ended and constructed response items (Hong & Wilcox, in preparation)
- Longitudinal topic modeling
    - Topic measurement invariance
- Exploratory vs. confirmatory topics and validation

Questions?

kwilcox3@nd.edu

Wilcox, K. T., Jacobucci, R., Zhang, Z., & Ammerman, B. A. (under review). Supervised latent Dirichlet allocation with covariates: A Bayesian structural and measurement model of text and covariates. *PsyArXiv*. doi: 10.31234/osf.io/62tc3